SALICON: Saliency in Context

Ming Jiang† Shengsheng Huang† Juanyong Duan† Qi Zhao*

Department of Electrical and Computer Engineering, National University of Singapore

mjiang@nus.edu.sg shane.huang@nus.edu.sg j.duan@u.nus.edu eleqiz@nus.edu.sg

Abstract

Saliency in Context (SALICON) is an ongoing effort that aims at understanding and predicting visual attention. This paper presents a new method to collect large-scale human data during natural explorations on images. While current datasets present a rich set of images and task-specific annotations such as category labels and object segments, this work focuses on recording and logging how humans shift their attention during visual exploration. The goal is to offer new possibilities to (1) complement task-specific annotations to advance the ultimate goal in visual understanding, and (2) understand visual attention and learn saliency models, all with human attentional data at a much larger scale.

We designed a mouse-contingent multi-resolutional paradigm based on neurophysiological and psychophysical studies of peripheral vision, to simulate the natural viewing behavior of humans. The new paradigm allowed using a general-purpose mouse instead of an eye tracker to record viewing behaviors, thus enabling large-scale data collection. The paradigm was validated with controlled laboratory as well as large-scale online data. We report in this paper a proof-of-concept SALICON dataset of human "free-viewing" data on 10,000 images from the Microsoft COCO (MS COCO) dataset with rich contextual information. We evaluated the use of the collected data in the context of saliency prediction, and demonstrated them a good source as ground truth for the evaluation of saliency algorithms.

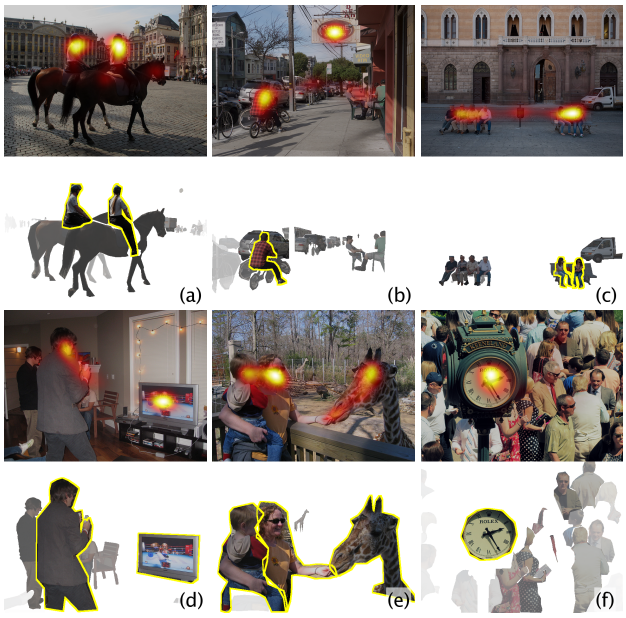

Figure 1. Contextual information is crucial in image understanding (image examples from MS COCO). We propose a new method to collect large-scale attentional data (SALICON, 1st row) for in visual understanding. With the annotated object segments, our attentional data naturally highlights key components in an image (ranked object segments in the 2nd row, with key objects outlined in yellow) to (a) rank object categories, (b) suggest new categories important to characterize a scene (text in this example), (c-e) convey social cues, and (f) direct to places designed for attention in advertisement.

1. Introduction

Motivation One of the ultimate goals in computer vision is to describe the contents of an image. Humans are known to perform better than their machine counterparts in telling a story from an image, and we aim to leverage human intelligence and computer vision algorithms to bridge the gap between humans and computers in visual understanding.

In the recent years, several datasets have been constructed with unprecedented numbers of images and annotations [32, 6, 34, 19], enabling breakthroughs in visual scene understanding, especially goal-specific tasks like object classification and segmentation. In the recently published MS COCO dataset [19], non-iconic images and objects in context are emphasized to understand natural scenes. On top of annotations for the conventional computer vision

†The three authors contribute equally to this work.
*Corresponding author.

tasks, it also includes sentences to describe an image, a big step toward the Turing test in the visual domain.

Complementary to all the existing big datasets, in this work we focus on how people direct their gaze when inspecting a visual scene. Humans and other primates shift their gaze to allocate processing resources to the most important subset of the visual input. Understanding and emulating the way that human observers free-view a natural scene to respond rapidly and adaptively has both scientific and economic impact. The logging of human viewing data during the assumption-free exploration also provides insights to other vision tasks and complement them to better understand and describe image contents (see Figure 1). For example, it naturally ranks labeled object categories, and suggests new categories for current classification datasets. By highlighting important objects by humans, it leverages human intelligence in visual understanding.

To collect large-scale human behavioral data in scene exploration, we first propose a novel psychophysical paradigm to record mouse-movement data that mimic the ways humans view a scene [31]. The designed stimuli encode the visual acuity drop-off as a function of retinal eccentricity. The mouse-contingent paradigm motivates mouse movements, to reveal interesting objects in the periphery with high resolution, similarly as humans shift their gazes to bring objects-of-interest to the fovea. Rather than recording the task-specific end outcomes by human annotators, we record the natural viewing patterns during the exploration. Therefore, our method is general and task-free. We then propose a crowdsoucing mechanism to collect large-scale mouse-tracking data through Amazon Mechanic Turk (AMT).

Challenges To record where humans look, eye-tracking experiments are commonly conducted, where subjects sit still in front of a screen with their eye movements recorded by a camera. Normally an infrared illuminator is necessary to help acquire high-quality data. There are several challenges particular to data collecting and usage.

First, large-scale data collection is prohibitive. An eye tracker used in laboratories generally costs between $30,000 - $80,000. Despite recent advances in gaze and eye modeling and detection (e.g., [10]), accurate eye-tracking experiments are still difficult without customized eye-tracking hardware. Data collection with general-purpose webcams is not yet possible, especially in uncontrolled settings such as through the AMT platform. This greatly limits the data collection process. As a result, the sizes of the current eye-tracking datasets are at the order of hundreds images and tens subjects, much smaller than those for object detection, object classification, scene categorization, or segmentation.

Second, eye-tracking data are not sufficiently general. Datasets collected from different labs are quite different in nature due to various image selection criteria, experimental setup, and instructions. Thus datasets cannot be directly combined, nor models learned from one dataset directly generalize to another [38].

Objectives This paper focuses on two major objectives:

1. We propose a novel psychophysical paradigm as an alternative to eye tracking, to provide approximation of human gaze in natural exploration. We design a gaze-contingent multi-resolutional mechanism where subjects can move the mouse to direct the high-resolutional fovea to where they find interesting in the image stimuli. The mouse trajectories from multiple subjects are aggregated to indicate where people look most in the images.
2. We propose a crowdsourcing platform to collect large-scale mouse-tracking data. We first sample 10,000 images from the MS COCO dataset with rich contextual information, and collect mouse-movement data using AMT. The "free-viewing" dataset is by far the largest one in both scale and context variability. We would like to point out that, with the crowdsourcing platform, it allows us to easily collect and compare various data with different top-down instructions, for example, to investigate the attention shifts during story-telling vs. category labeling.

2. Related work

Eye-tracking datasets There is a growing interest in the cognitive science and computer science disciplines to understand how humans and other animals shift their gazes to interact with the complex visual scenes. Several eye-tracking datasets have been recently constructed and shared in the community to understand visual attention and to build computational saliency models.

An eye-tracking dataset includes natural images as the visual stimuli and eye movement data recorded using eye-tracking devices. A typical dataset contains hundreds or a thousand images, viewed by tens of subjects while the locations of their eyes in image coordinates are tracked over time. Even if POET, the largest dataset we know by far, contains 6,270 images and is only viewed by 5 subjects [21]. While instructions are known to affect eye movement patterns, most common in eye-tracking dataset is the use of a so-called "free-viewing" task [5, 18, 4, 27] due to its task-free nature.

Most datasets have their own distinguishing features in image selection. For example, most images in the FIFA dataset [5] contain faces, and the NUSEF dataset [27] focuses on semantically affective objects/scenes. Compared with FIFA and NUSEF, the widely used Toronto dataset has less noticeably salient objects in the scenes. The MIT dataset [18] is more general due to its relatively large size, *i.e.*, 1003 images, and the generality of the image source. Quite a few images in these datasets are with dominant objects in the center. To facilitate object and semantic saliency,

the OSIE dataset [35] features in multiple dominant objects in an image. Besides general purpose images, there are also recent datasets in focused domains like the MIT Low Resolution dataset [17] for saliency in low resolution, EyeCrowd [16] for saliency in crowd, and FiWI [28] for web page saliency. Human labeling such as object bounding boxes [16], contours [27, 35], and social attributes [35, 16] are available in certain datasets as ground truth data for learning and analysis of problems of interest.

The scale of the current datasets is inherently limited by the experimental requirements. We envision that the collection of a larger-scale eye-tracking dataset would not only improve saliency prediction with big ground truth data, but driving new research directions in visual attention studies as well as complementing current efforts in computer vision datasets and annotations for more ambitious tasks in visual understanding.

Crowdsourcing Manual labeling to obtain ground truth human data is important for computer vision applications. Human knowledge and experience in this way is leveraged to train better computer models. Services like Amazon Mechanical Turk (AMT) has been extensively used to distribute the labeling task to many people, allowing the collection of large-scale labeling data. Recent works [32, 6, 33, 34, 7, 19] mainly focused on crowdsourcing image classification, object detection, and segmentation using this marketplace. Some of the most successful datasets along the line include Tiny Images [32], ImageNet [6] SUN [34], and MS COCO [19]. These datasets include hundreds thousands to millions of images containing hundreds or thousands of categories of interest, aiming at capturing general objects, scenes, or context in the visual world.

Current crowdsourcing tasks focus on the end output from humans (*e.g.*, a category label, an object segment), while our method records the procedure during which humans explore the scene in a real-time manner. We expect that the viewing patterns reveal cognitive process and can be leveraged for intelligent visual understanding. Our current experiments use task-free scenarios, and it could work with any other task-specific annotation procedure to log how humans explore the scene to complete a certain task.

Mouse tracking Mouse tracking and eye-mouse coordination have been studied in the human-computer interaction literature. For example, one of the most popular application of mouse-tracking is web page analysis [13, 20]. Huang *et al.* [13] studied mouse behaviors in web searching tasks, suggesting the plausibility of using mouse positions to predict user behavior and gaze positions. Navalpakkam *et al.* [20] integrated the mouse position on web pages with task relevance, and developed computational models to predict eye movement from mouse activity. Web pages contain domain-specific contents that motivate users to move their mouse to click links and to navigate. In natural images, however, to motivate users to move their mouse as one shifts attention requires specific design of the visual stimuli.

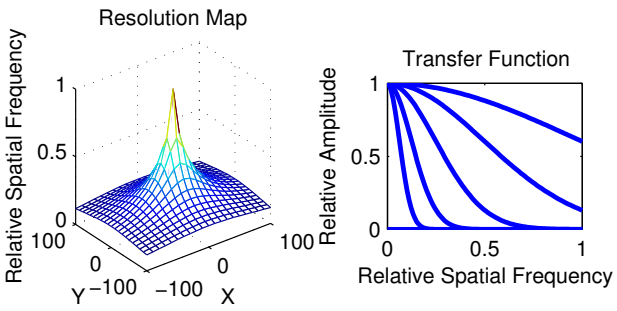

Figure 2. The resolution map and transfer functions.

3. Mouse-contingent free-viewing paradigm

To verify the feasibility of replacing eye-tracking data collection with mouse tracking, and to investigate the correlations between the two modalities, we designed a novel mouse-contingent paradigm with multi-resolutional images generated in real-time. We compared mouse-tracking data with eye-tracking data on the OSIE dataset that contains 700 images with the resolution of 800×600. The mouse-tracking data were collected in a controlled laboratory environment, with similar hardware and software configurations as reported in [35].

3.1. Stimuli

To simulate the free-viewing patterns of human visual attention with mouse tracking, we created an interactive paradigm by producing multi-resolutional images in real-time, based on the simulation method proposed by Perry and Geisler [24]. Gaze-contingent and mouse-contingent stimuli have been used in a variety of psychophysical studies, such as reading [1] and visual search [26]. The production of multi-resolutional images is based on neurophysiological and psychophysical studies of peripheral vision. Human visual system shows a well-defined contrast sensitivity by retinal eccentricity relationship. Specifically, contrast sensitivity to higher spatial frequencies drops off as a function of retinal eccentricity (e.g., [22, 25]). Therefore, we first generated a resolution map to simulate the sensitivity drop-off in peripheral vision [14] (see Figure 2). It is defined as a function $R : \Theta \rightarrow [0, 1]$, where Θ is the set of viewing angles θ with respect to the retinal eccentricity, and $[0, 1]$ represents the set of relative spatial frequency. The resolution map approximates a normal adult's vision with the exclusion of the blind spot. A higher $R(\theta)$ indicates a higher resolution at the visual eccentricity θ. Specifically, the resolution map is formulated as

$$R(x, y) = \frac{\alpha}{\alpha + \theta(x, y)}, \tag{1}$$

Figure 3. An example of the mouse-contingent stimuli. The red circles indicate the movement of mouse cursor from one object to another.

where $\alpha = 2.5°$ is the half-height angle, which means that when $\theta(x, y) = \alpha$ the image will become only half the resolution of the center of attention ($\theta(x, y) = 0$). In our experiments, we set $\alpha = 2.5$ to approximate the actual acuity of human retina. The image coordinates were mapped to the visual angles by the following function:

$$\theta(x, y) = \frac{1}{p}\sqrt{(x - x_c)^2 + (y - y_c)^2}, \tag{2}$$

where θ is the visual angle, x and y are pixel coordinates, and (x_c, y_c) is the center of attention. The parameter p represents the number of pixels a person can see in a degree of visual angle, which can be changed to simulate different viewing distances. Generally, the closer the distance is, the less can be seen in the high-resolutional fovea. We found that $p = 7.5$ led to a more comfortable and natural experience, according to the subjects' performances and feedbacks in pilot experiments. An example of the produced multi-resolutional image is shown in Figure 3. To compute the multi-resolutional image in real-time, we applied a fast approximation with a 6-level Gaussian pyramid from A_1 to A_6. A_1 was the original image and A_i was down-sampled to A_{i+1} with a factor of 2 in both dimensions. The standard deviation of the Gaussian distribution was set to $\sigma = 0.248$ pixel. After that, all the down-sampled images (A_2 to A_6) were then interpolated to the original image size. We then computed six matrices of blending coefficients, $M_1 \cdots M_6$. We used transfer function $T(f)$ (see Function 3 and Figure 2) and blending function $B(x, y)$ (see Equation 1 in [24]) to calculate these blending coefficients. The transfer function maps relative spatial frequency $f = R(x, y)$ to relative amplitude $T(f)$ in the Gaussian pyramid:

$$T_i(f) = \begin{cases} e^{1/2\times(-2^{i-3}f/\sigma)^2}, & i = 1, \cdots, 5 \\ 0, & i = 6, \end{cases} \tag{3}$$

The blending function $B(x, y)$ calculates the blending coefficients of each pixel (x, y):

$$B(x, y) = \frac{0.5 - T_i(x, y)}{T_{i-1}(x, y) - T_i(x, y)}, \tag{4}$$

where i is the layer number of (x, y). To calculate the layer number, we first determined six bandwidths $w_i, i = 1 \cdots 6$ such that $T_i(w_i) = 0.5, i = 1 \cdots 5$ and $w_6 = 0$. Then we normalized all w_i to $[0, 1]$. The layer number of pixel (x, y) is i such that $w_{i-1} \geq R(x, y) \geq w_i$. Next we calculated entries of $M_1 \cdots M_6$. For each pair of indices (x, y), we considered it as a pair of coordinates of a pixel and we calculated its layer number i_0, then

$$M_i(x, y) = \begin{cases} B(x, y), & i = i_0 - 1 \\ 1 - B(x, y), & i = i_0 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

for $i = 1 \cdots 6$. Finally, the multi-resolutional stimulus was a linear combination of M_i and A_i for $i = 1 \cdots 6$.

3.2. Subjects and procedure

Sixteen subjects (10 male and 6 female) aged between 19 and 28 participated in the mouse-tracking experiment. All participants had normal or corrected-to-normal vision, and normal color vision as assessed by Ishihara plates. All subjects had not participated in any eye-tracking experiment or seen the OSIE images before. The images were presented to the subjects in 700 trials at random order. Each trial consists of a 5-second image presentation followed by a 2-second waiting interval. The mouse cursor was displayed as a red circle with a radius of 2 degrees of visual field that is sufficiently large not to block the high-resolutional region of focus, and automatically moved to the image center when the image onset. The subjects were instructed to explore the image freely by moving the mouse cursor to anywhere they wanted to look. No further instructions were given on how to move the mouse or where they should look in the images. Whenever they moved the mouse, the mouse-contingent stimuli was updated by shifting the center of the resolution map to the mouse position. In the meantime, the mouse position and the timestamp were recorded. Each block contains 50 trials, and the subject can take a short break between blocks.

Presentation of stimuli and recording of mouse position were implemented in Matlab (Version 8.1.0, Mathworks, MA) using the Psychophysics Toolbox [2, 23]. The experiment PC was a Dell T5610 (2.5GHz, 32GB RAM, Ubuntu 14.04) with a Quadro K600 graphics card. The mouse speed and acceleration were adjusted to the maximum in the system settings. There was a practice session for the subjects to get familiar with the mouse-contingent paradigm and the mouse configuration, which consists of 10 other images from the Internet with the same resolution as the OSIE images. The practice trials were identical to the formal trials in terms of all parameters.

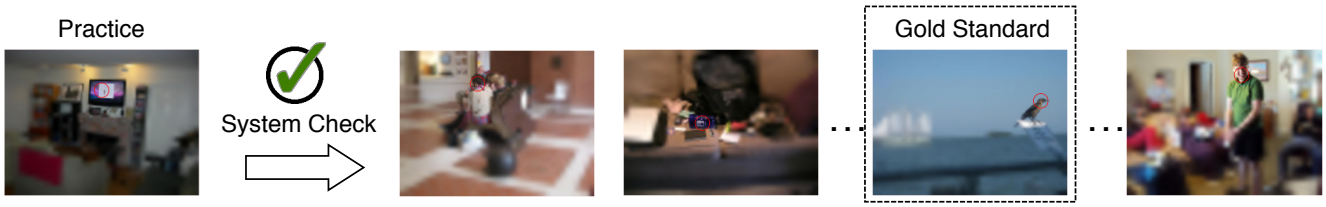

Figure 4. The procedure of an AMT task.

4. Large-scale attentional data collection by AMT deployment

The motivation for the mouse-tracking paradigm is for large-scale data collection. In this section, we report implementation and design issues to deploy the mouse-tracking experiments on the paid AMT crowdsourcing marketplace. We employed the same paradigm and parameter configurations as described in Section 3, while making a few minor adjustments to the procedure to accommodate the more uncontrolled online situations. Figure 4 illustrates the online experiment procedure on AMT.

Our task required real-time rendering of the mouse-contingent stimuli, *i.e.*, the image rendering was triggered by the mouse events in the browser. Therefore it was important to perform a system check to ensure a smooth rendering during visual exploration. The system check was conducted at the practice stage of an AMT task, which detected failures due to a variety of reasons such as unsupported browser features, unfriendly browser plug-ins, and low memory capacity. To ensure that our paradigm was shown smoothly without noticeable lag at the browser side, we evaluated the synchronization quality of the display and the mouse activity, by measuring the distances between the mouse positions and the rendered centers of attention. Only workers who passed the system check could continue the task.

We deployed the experiment on AMT using 10,000 MS COCO training images with 640×480 pixels and 700 OSIE images (scaled to 640×480 pixels). The OSIE images were added as "gold standard", where the eye-tracking data in OSIE can be used as a baseline to evaluate the performance of workers. Currently in each task, a worker viewed 40 images, including 36 images from the MS COCO dataset and 4 images from the OSIE dataset. With the large-scale data collection, we created a Saliency in Context (SALICON) dataset, with 10,000 MS COCO images viewed by 60 observers each. Details of the mouse-tracking results and statistics of the experiments are reported in Section 5.

5. Statistics and results

In this section, we report the mouse-tracking statistics of the two datasets – OSIE and SALICON. For OSIE images, we compare three sets of data: eye tracking, mouse tracking in lab, and mouse tracking with AMT. For SALICON, we report the mouse-tracking statistics in terms of the MS COCO object categories.

5.1. Data preprocessing

Due to the differences in hardware and software settings, the mouse-tracking data have a large variety of sample rates. In the lab experiments, the mean sample rate was 285.61 Hz, across all subjects. While in the AMT data, due to the event system of the browser environments, the sampling was not triggered until the mouse moved. Therefore, the mean sample rate was 69.42 Hz. We discarded the data with sample rate lower than 12 Hz, and resampled the data with a shape-preserving piecewise cubic interpolation that matched the data in position, velocity and acceleration. This was to equalize the number of samples across all observers. The normalized mouse samples had a uniform sample rate at 100 Hz. We added a simple pre-processing step by excluding half samples with high mouse-moving velocity (*i.e.* saccades) for each observer, while keeping the fixations. All pre-processed mouse samples for the same image were then aggregated and blurred with a Gaussian filter to generate a saliency map, same as the common practice to generate the fixation maps from eye-tracking data [35].

5.2. Center bias

In almost all eye-tracking datasets, there exists a spatial prior that pixels near the image center attract more fixations, known as the center bias [30]. The main reasons of the center bias include photographer bias, experimental configuration, and viewing strategy. Similarly, our mouse-tracking data are also biased towards the image center. The cumulative distribution of the mean distance from sample points to the image center is shown in Figure 5. We normalized the distance to center by the image width, and did not observe significant differences in the average distance to center between the AMT and controlled mouse-tracking data or between mouse-tracking and eye-tracking data.

5.3. Evaluating mouse maps with eye fixations

We evaluated the similarities of the mouse maps and the eye fixation maps, using the most commonly used evaluation metric – the shuffled AUC (sAUC) [37]. The sAUC computes the area under the receiver operating characteristic (ROC) curve, taking positive samples from the fixations

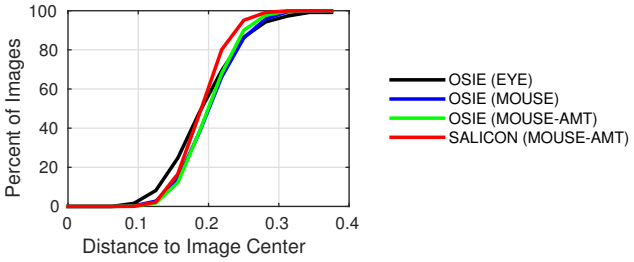

Figure 5. Cumulative distributions of mean distances from fixations / samples to the image center. Distances are normalized with the image width.

of a test image, and negative samples from all fixations in other images. This way it discounts the global center bias in the dataset. We compared the performances of the mouse maps with the inter-observer performance of eye tracking (computed by aggregating fixations from other subjects than each tested subject, used as a baseline). We also included the highly referred and the state-of-the-art saliency algorithms in the comparison [15, 11, 37, 3, 9, 12, 36]. All saliency maps were blurred by a Gaussian kernel with σ from 0 (no blurring) to 3 degrees of visual angle (DVA; 24 pixels according to the eye-tracking configuration), and the optimal blur width was chosen for each model.

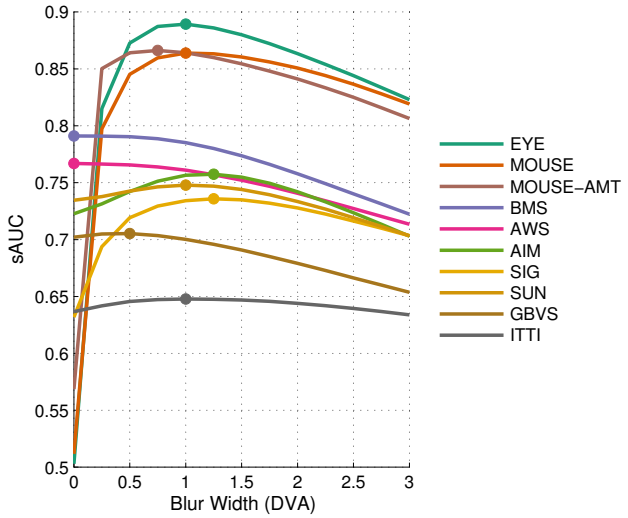

Figure 6. Eye fixation prediction performance with mouse tracking and the highly referred/state-of-the-art computational saliency models: eye tracking (EYE), mouse map in lab (MOUSE), mouse map on AMT (MOUSE-AMT) the Itti & Koch model (ITTI) [15], the information maximization model (AIM) [3], the graph-based saliency (GBVS) [11], the saliency using natural statistics (SUN) [37], the image signature (SIG) [12], the adaptive whitening saliency (AWS) [9], and the boolean map saliency (BMS) [36].

As shown in Figure 6, the lab and AMT mouse models scored closely in sAUC ($\sim$ 0.86). They are much closer to

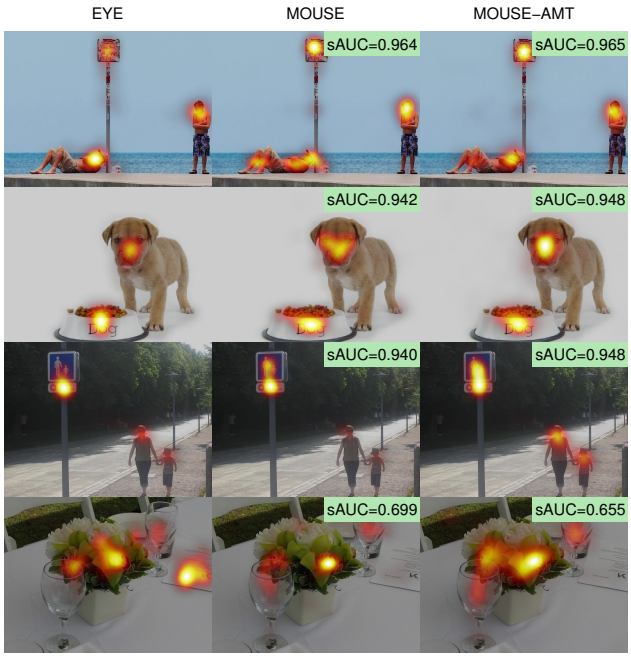

Figure 7. Image examples with high and low eye-mouse similarities evaluated with sAUC. Eye fixation maps and mouse maps are overlaid.

the human performance ($\sim$ 0.89) in eye tracking than the computational models. Figure 7 presents the images with high and low sAUC scores in mouse tracking (with AMT). While the mouse-eye agreement is high in simple images, it is generally lower in more complex scenes, where inter-observer consistency in the eye-tracking data is also lower (Pearson's correlation on sAUC r=0.76, p$<$0.001). Further, mouse tends to miss far and indistinguishable text, not only because mouse is slower than eye [29], but also due to the the relatively low peripheral resolution of text [17]. This may be caused by the relatively small visual angle we use (7.5 pixels per degree) in the mouse-contingent paradigm. As described in Section 3, the free parameter p corresponds to the visual angle to the scene, ecologically valid in natural vision. While the conventional eye-tracking experiments mostly fix this parameter, the proposed paradigm allows the change of this parameter to mimic scenarios with varying distances to the stimuli.

5.4. Categorical analysis

For the SALICON dataset, we sampled 10,000 images from the currently released MS COCO training set, which contains 80 of the 91 categories. The subset was selected from a total of 17,797 images with the resolution of 640$\times$480. The selection was based on the number of categories in each image. Figure 8 reports the statistics of the dataset in comparison with the MS COCO training images. Our selected images have more instances and categories per

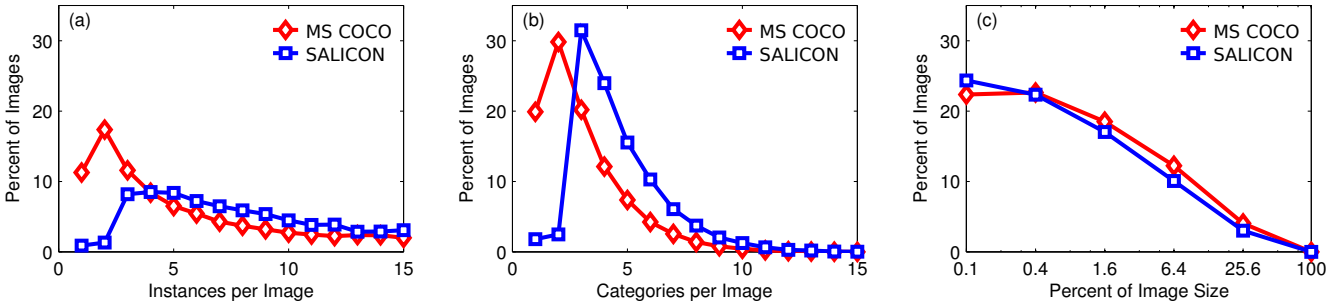

Figure 8. Distributions of (a) number of instances per image, (b) number of categories per image, and (c) instance area.

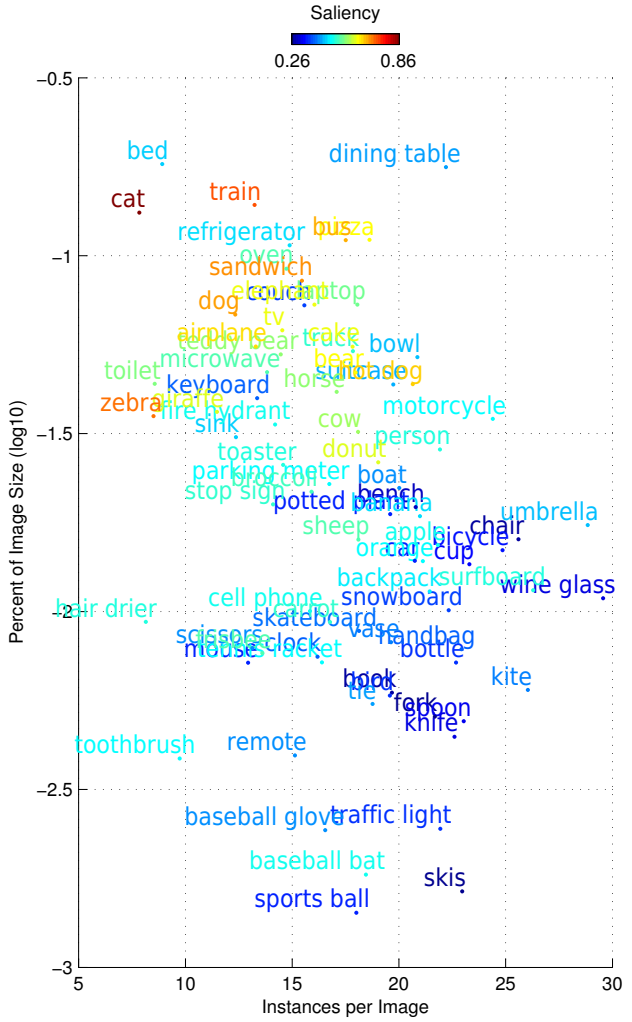

Figure 9. Average saliency values for each of the 80 object categories in SALICON.

image, and are in general richer in contextual information. The rich context is helpful to compare the relative importance of each category in visual exploration. The instance sizes in the SALICON is not significantly different from the full MS COCO training set.

With the mouse maps from the aggregated AMT data, we computed the "maximum object saliency" as the maximum of the map values inside each object's outline, as it does not scale with the object size [8]. This way we rank the objects in the same image by these values to decide their relative importance.

To quantify the importance of categories with the attentional data, for each category of instances, we computed the mean instance size, the average number of total instances in the scene which has instances of the particular category, and average saliency value. Figure 9 shows the average saliency values for all the 80 object categories in our dataset. As observed, the importance of a category correlates with its average size and number of instances in the same scene. For the most salient categories, objects appear relatively large in images and are with fewer distracters. Examples include animals, food and train. In comparison, furniture like bed, dining table and refrigerator are relatively less salient, although large in size. Small objects are mostly less salient, except categories that are interactive with humans such as surfboard, baseball bat, and tennis racket.

We further explored the collected attentional data as a natural way to suggest new categories for object annotations and segmentation. The MS COCO has selected 91 categories leveraging domain references, children experiences, and mutual agreement from co-authors. Human attentional data provide yet another complementary source that identify objects that humans look at frequently and rapidly during natural exploration. Figure 10 illustrates examples of typical scenarios where fixations land on unlabeled objects, and suggests several categories be added to the MS COCO dataset to improve its contextual richness. For example, faces attract attention consistently and strongly. Since it is not defined as a category but subregion of 'person', we observe that (1) most fixations land on faces though the entire persons are annotated, and (2) some faces are missed if the objects do not belong to the existing category (*e.g.*, toy face, animal face in the first row in Figure 10). Text and pictures also attract attention consistently, but not explicitly defined category in MS COCO (second and third rows). As illustrated in the fourth row, food is frequently missed as

Figure 10. Examples of salient but missed object categories, including face, text, picture, food, door and window, etc. Segmented object instances are masked with colors indicating the categories.

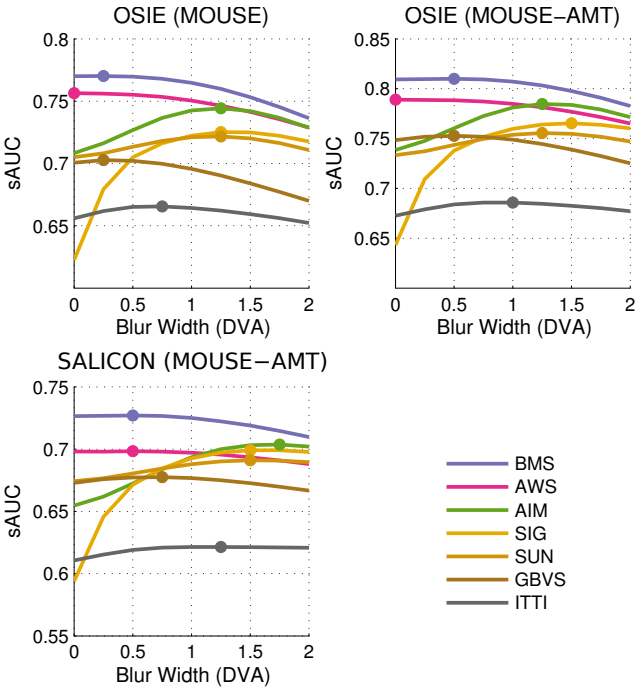

Figure 11. Evaluation of saliency algorithms against mouse-tracking data.

only certain types of food are defined (*e.g.*, broccoli, sandwich). Doors and windows attract considerable gaze (fifth row) mostly due to their contextual importance. Detecting these objects would help to understand the context of the scene. These examples demonstrate a potential application of the proposed work in complementing other annotations for visual understanding.

5.5. Mouse tracking as an evaluation benchmark

Since the mouse-tracking and eye-tracking data were qualitatively and quantitatively similar, we further exploited the mouse tracking as a benchmark to evaluate computational saliency algorithms. We tested the state-of-the-art saliency algorithms on the OSIE dataset and randomly selected 2,000 images from the SALICON dataset. We used the pre-processed mouse samples as positive samples in the sAUC computation. For the AMT mouse-tracking data (OSIE and SALICON), in order to reduce the computational cost in the evaluation, we filtered the mouse samples by only keeping the pixels viewed by at least two observers. The comparative results are shown in Figure 11. From the comparison we observe that on OSIE, the sAUC scores for both mouse-tracking data (laboratory and AMT) are close to the eye-tracking ones (see Figure 6), and their ranks are basically preserved. The results show that mouse tracking is a good replacement of eye tracking in model evaluation. Comparing the saliency algorithm performance on SALICON vs. on OSIE, similar patterns are observed too. The difference in score reflects dataset difference in image properties.

6. Conclusion

This paper presents a new paradigm to collect human attentional data. Our paradigm enables large-scale data collection by using a general-purpose mouse instead of an expensive eye tracker to record viewing behaviors. With the proposed method, a large mouse-tracking dataset for saliency in context (SALICON) was created on 10,000 images from MS COCO. SALICON is by far the largest attention dataset in both scale and context variability, and data collection on more images is ongoing with the same protocol. With the visual attentional data collected from mouse tracking, the SALICON dataset complements existing task-specific annotations with natural behavior of visual exploration in task-free situations. The paradigm can also be easily generalized to various types of tasks with top-down instructions. We also envision SALICON to be a good source for learning and benchmarking saliency algorithms with more data.

Acknowlegements

The research was supported by the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO, the Defense Innovative Research Programme (No. 9014100596), and the Ministry of Education Academic Research Fund Tier 1 (No. R-263-000-A49-112).

References

[1] What guides a reader's eye movements? *Vision Res.*, 16(8):829–837, 1976. 3

[2] The psychophysics toolbox. *Spat. Vis.*, 10(4):433–436, 1997. 4

[3] N. Bruce and J. Tsotsos. Saliency based on information maximization. In *NIPS*, pages 155–162, 2005. 6

[4] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: an information theoretic approach. *J. Vis.*, 9(3):5.1–24, 2009. 2

[5] M. Cerf, E. P. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *J. Vis.*, 9(12):10.1–15, 2009. 2

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 3

[7] J. Deng, J. Krause, and F.-F. Li. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, pages 580–587, 2013. 3

[8] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. Vis.*, 8(14):18.1–26, 2008. 7

[9] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image Vis. Comput.*, 30(1):51–64, 2012. 6

[10] D. W. Hansen and Q. Ji. In the eye of the beholder: a survey of models for eyes and gaze. *TPAMI*, 32(3):478–500, 2010. 2

[11] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *NIPS*, pages 545–552, 2006. 6

[12] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *TPAMI*, 34:194–201, 2012. 6

[13] J. Huang, R. W. White, G. Buscher, and K. Wang. Improving searcher models using mouse cursor activity. In *SIGIR*, page 195, 2012. 3

[14] H.-W. Hunziker. *Im Auge des Lesers: foveale und periphere Wahrnehmung - vom Buchstabieren zur Lesefreude*. Transmedia Verlag, 2006. 3

[15] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11):1254–1259, 1998. 6

[16] M. Jiang, J. Xu, and Q. Zhao. Saliency in crowd. In *ECCV*, 2014. 3

[17] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *J. Vis.*, 11(4):14.1–20, 2011. 3, 6

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009. 2

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 1, 3

[20] V. Navalpakkam, L. L. Jentzsch, R. Sayres, S. Ravi, A. Ahmed, and A. Smola. Measurement and modeling of eye-mouse behavior in the presence of nonlinear page layouts. In *ICWWW*, pages 953–964, 2013. 3

[21] D. P. Papadopoulos, A. D. F. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *ECCV*, pages 361–376, 2014. 2

[22] E. Peli, J. Yang, and R. B. Goldstein. Image invariance with changes in size: the role of peripheral contrast thresholds. *J. Opt. Soc. Am. A*, 8(11):1762, 1991. 3

[23] D. G. Pelli. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.*, 10:437–442, 1997. 4

[24] J. S. Perry and W. S. Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *SPIE*, pages 57–69, 2002. 3, 4

[25] J. S. Pointer and R. F. Hess. The contrast sensitivity gradient across the human visual field: with emphasis on the low spatial frequency range. *Vision Res.*, 29(9):1133–1151, 1989. 3

[26] M. Pomplun, E. M. Reingold, and J. Shen. Investigating the visual span in comparative search: The effects of task difficulty and divided attention. *Cognition*, 81(2):B57–B67, 2001. 3

[27] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T. S. Chua. An eye fixation database for saliency detection in images. In *ECCV*, pages 30–43, 2010. 2, 3

[28] C. Shen and Q. Zhao. Webpage saliency. In *ECCV*, pages 33–46, 2014. 3

[29] L. E. Sibert and R. J. Jacob. Evaluation of eye gaze interaction. In *SIGCHI*, pages 281–288, 2000. 6

[30] B. W. Tatler. The central fixation bias in scene viewing: selecting an optimal viewing position independently of motor biases and image feature distributions. *J. Vis.*, 7(14):4.1–17, 2007. 5

[31] L. N. Thibos. Acuity perimetry and the sampling theory of visual resolution. *Optom. Vis. Sci.*, 75(6):399–406, 1998. 2

[32] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 30(11):1958–1970, 2008. 1, 3

[33] P. Welinder, S. Branson, P. Perona, and S. Belongie. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432, 2010. 3

[34] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 1, 3

[35] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *J. Vis.*, 14(1):28.1–20, 2014. 3, 5

[36] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *ICCV*, pages 153–160, 2013. 6

[37] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *J. Vis.*, 8(7):32.1–20, 2008. 5, 6

[38] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *J. Vis.*, 11(3):9.1–15, 2011. 2